Text

Description automatically generated

STATISTICS FOR HIGH DIMENSIONAL DATA

**PROJECT REPORT**

**Classification of cardiac arrhythmias**

Mously DIAW – M2 IMSD

**Academic year 2020-2021**

Table of contents

[List of figures 3](#_Toc56952045)

[List of tables 3](#_Toc56952046)

[1. Introduction 4](#_Toc56952047)

[2. Data preprocessing and exploratory data analysis 4](#_Toc56952048)

[3. Data splitting and model strategy 6](#_Toc56952049)

[4. Results 7](#_Toc56952050)

[5. Final model 8](#_Toc56952051)

[6. References 9](#_Toc56952052)

[7. Annexes 10](#_Toc56952053)

# List of figures

**No table of figures entries found.**

# List of tables

**No table of figures entries found.**

# Introduction

Heart arrhythmias are characterized by irregular heartbeats, which could also be too slow or too fast. In order to diagnose a cardiac arrhythmia, the heart activity is analyzed by recording an electrocardiogram (ECG). The parameters of the ECG combined with patient information allows to detect ad categorize arrhythmia. Unfortunately, Intensive care units (ICU) false arrhythmia alarm rates have been reported to be as high as 88.8% [1]. This high percentage negatively impacts both patients and clinical staff. The alarm overload might also lead to desensitization and could result in true alarms being ignored.

Guvenir et al This database arrhytmia dataser that can be found in the UCl Machine Learning Repository [2] . aimed at solving this issue by distinguish between the presence and absence of cardiac arrhythmia and to classify it in one of the 16 groups. Class 01 refers to 'normal' ECG classes 02 to 15 refers to different classes of arrhythmia and class 16 refers to the rest of unclassified ones.

In this project, I will use the arrhythmia dataset to create a classification model that will distinguish between the normal and anormal arrhythmia. This approach is due to the severe class imbalance noticed in the dataset. Before selecting the best classifier and evaluating its performance, I conducted the following: data pre-processing and EDA, feature selection and model tuning. The results obtained at each step are detailed in the next sections.

# Data preprocessing and exploratory data analysis

The original dataset contains 279 attributes-452 patients, 206 of which are linear valued and the rest are nominal. The population study is composed of 203 men and 249 women across all ages (0-83). After converting the class attribute into a binary variable (207 anormal -positive class vs 245 normal Fairly balanced. Figure 1 shows boxplot for the variables height (1.a) and weight (1.b). Outliers height above 600cm aberrat=nt . Outliers: z score>3 replaced by median => 2 values height. The **outliers** in weight are realistic so didn’t replace them.

Chart

Description automatically generated

Figure 1:

5 variables had **missing values**. I dropped the variable ‘J’ where 83% of the values were missing. For the remaining variables which had a percentage below 5% I performed kNN imputation. I also dropped the variables that had **zero-variance** or near-zero variance (when the ratio between the two most frequent values was above 20). Finally, the variables II and IO were dropped for their **high correlation** (>0.9) with …

Shape, rectangle

Description automatically generated

Figure 2:

The final dataset has 142 predictors, all numerical except for the sex variable.

Prior to modeling, one would expect that the arrhythmia class would be to the heart rate. Indeed, it is reported that in some cases the irregular rhythm of heartbeat could be either too slow 60 bpm (beats per minute) or too fast 100 bpm [3]. Figure 3 shows the boxplot of the heart rate in the cases of normal and anormal heart rhythms. As expected, we can see that for normal arrythmia the heart rate is within the range 60-100 bpm. For anormal arrhythmia the range is more widespread and goes beyond the normal range. However, there is not a clear-cut rule to distinguish between anormal and normal rhythm by just looking at the heart rate.

Chart, box and whisker chart

Description automatically generated

Figure 3

The QRS duration is another important parameter in diagnosing heart failure. Figure 4 plots the heart rate against the qrs duration with hue diagnosis. We can see that the normal class is contained within the anormal one, overlap. Therefore, there is not an obvious rule to classify and we can also guess that classifiers such as the linear SVM (Support Vector Machines) or the kNN will not work well on this data. Radial SVM could work but it will most likely have a lot of false negatives (anormal classified as normal).

Chart, scatter chart, bubble chart

Description automatically generated

Figure 4

# Data splitting and model strategy

There are 452 samples available; 70% was used for training the algorithms while the remainder was used to evaluate the final candidate models. Five repeats of 10-fold cross–validation were used to tune the models (50 resamples).

Rather than creating models that maximize the overall accuracy or Kappa statistics, the aim in this project is to maximize both classes make up 50% of your dataset, or both make up a sizable fraction, and you care about your performance in identifying each class equally, then you should use the AUC, which optimizes for both classes, positive and negative.

A series of models were fit to the training set. The tuning parameter combination associated with the highest average auc value was chosen for the final model and used in conjunction with the entire training set. Table 1 contains the models investigated.

Table 1

|  |  |  |
| --- | --- | --- |
| Models | Preprocessing | Tuning parameters |
| Stepwise Logistic Regression | Center-scale (CS) | - |
| PLS | CS | ncomp = 1:10 |
| LDA 1 | CS | - |
| LDA 2 | CS + PCA | - |
| Sparse LDA | CS | - |
| Linear SVM | CS | tunelength = 10 |
| Radial SVM | CS | tunelength = 10 |
| Polynomial SVM | CS | tunelength = 4 |
| kNN | CS | k= 1:2:101 |
| Random Forest | - | mtry = 1:15 |
| Single CART tree | - | tunelength = 10 |
| GBM | - | interaction.depth = 1:2:7  n.trees = 100:50:1000  shrinkage = 0.01,0.1  n.minobsinnode = 10 |

# Results

The model results are shown in Figure where box plots of the resampling estimates of the mean cost value are shown. The linear models, such as LDA and PLS, did not do well here. Feature selection did not help the linear discriminant model, but this may be due to that model’s inability to han- dle nonlinear class boundaries. FDA also showed poor performance in terms of cost.

There is a cluster of models with average costs that are likely to be equivalent, mostly SVMs and the various tree ensemble methods. Using costs/weights had significant positive effects on the single CART tree and SVMs. Figure 17.6 shows the resampling profiles for these two models in terms of their estimates of the cost, overall accuracy and Kappa statistic.

Trees clearly did well for these data, as did support vector machines and neural networks. Is there much of a difference between the top models? The p-values for the model comparisons are large (0.592 for accuracy and 0.269 for Kappa), which indicates that the models fail to show any difference in performance.

Chart, scatter chart

Description automatically generated

# Final model

Table 17.3 shows two such tables for the random forests model and the cost-sensitive CART model. For random forest, the average number of long jobs that were misclassified as very fast was 0.2 while the same value for the classification tree was 0.24. The CART tree shows very poor accuracy for the fast jobs compared to the random forest model. However, the opposite is true for moderately long jobs; random forest misclassified 72.56% of those jobs (on average), compared to 65.52% for the single tree. For long jobs, the single tree has a higher error rate than the ensemble method. How do these two models compare using the test set? The test set cost for random forest was 0.316 while the single classification trees had a average cost of 0.37. Table 17.4 shows the confusion matrices for the two models. The trends in the test set are very similar to the resampled estimates.

long jobs. In summary, the overall differences between the 2 models s not large.

# References

[1] Drew, Barbara J., et al. "Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients." *PloS one* 9.10 (2014): e110274.

[2] Guvenir, H. A., B. Acar, and H. Muderrisoglu. "Arrhythmia data set in UCI machine learning repository." *UC Irvine* (1998).

[3] Fu, Du-Guan. “Cardiac Arrhythmias: Diagnosis, Symptoms, and Treatments.” *Cell biochemistry and biophysics* vol. 73,2 (2015): 291-296.

# Annexes